I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

# Big Data and Open Data: How Open Will the Future Be?

JOEL GURIN

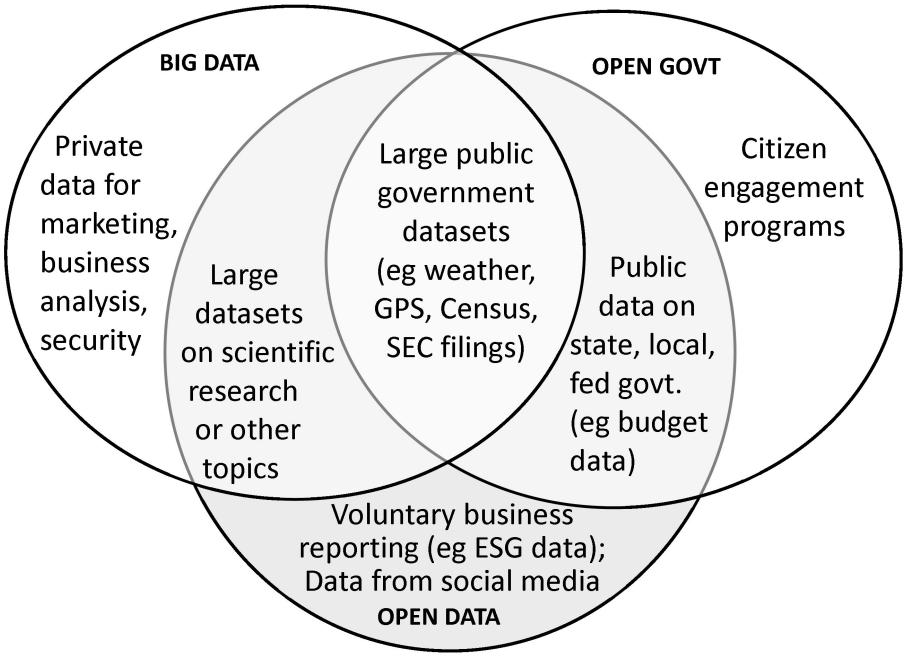
Big Data has become a phenomenon with impacts on government, business, science, and personal privacy. But it's only one of two major developments that are reshaping the relationships between citizens, society, and information. Equally important is the growth of Open Data, a related but very different concept.

Big Data describes datasets that are large, complex, and varied, that change rapidly, and that challenge the limits of our ability to analyze them. Open Data, in contrast, can be large or small, simple or complex. What distinguishes Open Data is the fact that it's been intentionally released for public use. I've described it as *accessible public data that people, companies, and organizations can use to launch new ventures, analyze patterns and trends, make data-driven decisions, and solve complex problems*.[1] The impact of data depends not only on how much data exists about a particular area, population, or individual, but how open the data is, who has access to it, and how it can be used to meet social or economic goals.

The Venn diagram below shows the relationship of Big Data and Open Data, and how the nature of a given dataset relates to how it is used. Section 1 of the diagram – Big Data that is *not* open – is the source of much of the public concern about data today. Data that the government collects about individuals for national security, or that retailers collect about their customers, gives these entities information that individuals may not want them to have. In contrast, Section 3 – Big Data that is *also* open – includes scientific data that can be shared between researchers to accelerate scientific progress and find new treatments and cures for serious illnesses. Finally, Section 6 – big,

[1] Joel Gurin, *Open Data Now* (New York: McGraw-Hill, 2014), 9.

open, government data – includes a wide range of important datasets that can lead to new social insights and help launch new businesses.




[1]

Big Data is most useful, and has the greatest economic and social value, when it is also Open Data. By releasing Big Data as Open Data, governments around the world can boost their countries' economies and improve the lives of their citizens. The intersection of Big Data and Open Data makes four major things possible: (1) using government data for business; (2) opening up data from scientific research; (3) using crowdsourcing to develop new kinds of Open Data; and (4) the use of data for smart cities.

At the same time, an "open" approach to data can help avoid many of the problems that Big Data might otherwise cause. An open approach can help make Big Data more useful; provide new approaches to dealing with privacy protection and personal data; and help create what might be called a "see-through society" with a healthy level of business and government transparency.

## I. BIG DATA MEETS OPEN DATA

### A. *Government Data for Business*

The federal government is now the most significant source of Big, Open Data in the U.S. Healthcare, education, energy, transportation, financial services, and other sectors are all being changed significantly by new Open Data ventures.

Transportation may be the most familiar example of making Big Data open and usable. Millions of people, for example, now use travel websites to book airline flights. Without these tools, it would be extraordinarily difficult to choose among the hundreds of flights each day between, say, Miami and Mexico City, and find the flight that will best meet one's own individual needs. These airline websites are examples of a particular kind of Open Data called Smart Disclosure – the use of data in "smart" applications to help consumers make complex choices.[2]

A nonprofit organization and website called GreatSchools[3] uses Open Data in this way to help parents find schools that will be the best fit for their children. GreatSchools uses both feedback that the site gets from parents and former students and data from formal statewide evaluations. This website is used by more than forty percent of all American households with K-12 children,[4] and demonstrates the interest that consumers have in advice and information derived from Open Data.

Some of the Open Data companies with the greatest impact go way beyond consumer services to develop applications that can affect an entire sector of the economy. Starting in 2006, the San Francisco-based Climate Corporation took data from the U.S. Weather Service, the National Oceanic and Atmospheric Administration, and other government sources, and hired brilliant data analysts and mathematicians who could extract new meaning from this public Open Data. The Climate Corporation started out to analyze the impact

---

[2] For more on Smart Disclosure, see National Science and Technology Council, "Smart Disclosure and Consumer Decision-Making," May 2013, http://www.whitehouse.gov/sites/default/files/microsites/ostp/report_of_the_task_force_on_smart_disclosure.pdf.

[3] "Great Schools," http://www.greatschools.org/.

[4] Joel Gurin, "Back to School with Open Data," *OpenDataNow.com* (blog), September 5, 2013, http://www.opendatanow.com/2013/09/back-to-school-with-open-data/.

of weather precisely in order to sell better weather insurance to farmers. But in order to do that, they had to develop a level of expertise and deep understanding of how weather, soil, rain, and agricultural conditions worked together to affect farming and the risks to crops.[5] In October 2013, Monsanto bought this young company for almost a billion dollars.[6] It's a dramatic example of how companies can get business value out of free data resources, adding value by working with the data in new, sophisticated ways.

Healthcare is likely to be the next frontier for big, open government data. An event called the Health Datapalooza,[7] which focuses on the uses of open health data in business, now attracts about two thousand people every year. One example of the new Open Data health companies is iTriage, which uses the open registry of healthcare providers around the country. iTriage allows a traveler in a strange city with chest pain to look up his or her symptoms, figure out if the problem needs urgent care, and, if so, find an emergency room nearby.[8] Another company, Aidin,[9] is using open government data to improve post-hospital care, while Evidera[10] uses Big Data on health to predict the outcomes of different treatments.

Open Data can help save energy. Opower,[11] a company based in the Washington, DC area, puts together data on energy efficiency with data about an individual household's own energy use, and then feeds that information back to consumers. They also provide an analysis of how much energy the neighbors use, in order to inspire people to save more energy than their neighbors do. Getting people to save energy is

---

[5] For an interview with the CEO of the Climate Corporation, see Gurin, *Open Data Now*, 27-31.

[6] Ashlee Vance, "Monsanto's Billion-Dollar Bet Brings Big Data to the Farm," *Bloomberg Businessweek Technology*, October 2, 2013, http://www.businessweek.com/articles/2013-10-02/monsanto-buys-climate-corporation-for-930-million-bringing-big-data-to-the-farm.

[7] "Healthdatapalooza," http://healthdatapalooza.org/.

[8] "iTriage," https://www.itriagehealth.com/.

[9] "Aidin," http://www.myaidin.com/.

[10] "Evidera," http://www.evidera.com/.

[11] "Opoer," http://opower.com/.

remarkably difficult, but Opower has had some significant success in promoting energy reduction.[12]

Finally, a number of companies are starting to make government data more useful to other companies. Enigma.io, for example, takes Open Data from federal sources, and increasingly from state and local sources, and puts them on a technical platform so that it can be used more easily for analysis and to generate new insights. In addition to making public data for free on their platform, Enigma.io makes more in-depth data available by subscription and offers enterprise services for companies that want to use Enigma's data for their particular needs.[13]

All of this raises an important question: What is open government data really worth? This is not just a theoretical concern, but also a public policy issue. Opening up data takes time, effort, and money, which ultimately come from taxpayers. It's important to know what the return will be.

Several researchers have estimated the potential economic value of Open Data. McKinsey and Company has estimated the value at $3 trillion per year worldwide, in a study that looked at data from both government and nongovernment sources.[14] While that number is significantly higher than other estimates,[15] open government data may well be worth tens or hundreds of billions of dollars a year. But these estimates have been based on high-level analysis and rely on a number of assumptions.

The Open Data 500 study,[16] now being done by the GovLab at NYU, is looking at the value of open government data from the ground up. The study is researching hundreds of U.S.-based companies that use open government data as key business resource. The Open Data

---

[12] Matt Davis, "Study Concludes Information-Based Energy Efficiency Can Save Americans Billions." *Environmental Defense Fund*, May 23, 2011, http://www.edf.org/news/study-concludes-information-based-energy-efficiency-can-save-americans-billions.

[13] "Enigma-Access the World's Public Data," http://enigma.io/plans/.

[14] James Manyika et al., "Open data: Unlocking Innovation and Performance with Liquid Information," McKinsey Global Institute, October 2013, http://www .mckinsey.com/insights/ business_technology/ open_data_unlocking_ innovation_and_performance_with_liquid_information.

[15] For a discussion of other estimates of the value of Open Data, see Joel Gurin, *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*, (New York: McGraw-Hill, 2014), ebook edition, 218-219.

[16] "Welcome to the Open Data 500," http://www.opendata500.com/.

500 is designed to (1) provide a basis for assessing the economic value of open government data, (2) encourage the development of new Open Data companies by showing examples of how Open Data can be used, and (3) identify the most valuable government datasets so agencies can prioritize their release. Building on the Open Data 500 study, the GovLab developed a series of Open Data Roundtables to bring together government agencies with the businesses and nonprofits that use their data to prioritize the most important goals.[17] By mid-2013, the Departments of Commerce, Labor, Transportation, and the Treasury, as well as the U.S. Department of Agriculture, had committed to participating in this process.[18]

### B. *Open Data for Scientific Research*

Big Data from scientific research is increasingly being released openly in a new model of data sharing. There are many incentives for researchers to keep data to themselves while work is under way. Scientists working in the pharmaceutical industry, for example, will generally keep data secret until they develop patentable products, while academic scientists want to keep data to themselves until they can publish it. But these secretive approaches aren't conducive to rapid progress. In contrast, they encourage competitive researchers to work in parallel, producing their own data, without benefiting from the kinds of collaborative approaches that would maximize the data's value.

A more collaborative model, and one that many scientists believe should be followed now, was exemplified by the Human Genome Project. In the mid-1990s, a few years after the project was launched, leading scientists met and agreed to share data with each other, an approach that accelerated their progress significantly.[19] In a similar way, several biomedical foundations, like the Multiple Myeloma Research Foundation,[20] are now requiring their grantees to share their data as a condition of funding.

---

[17] "Open Data Roundtables," http://www.opendata500.com/us/roundtables/.

[18] Joel Gurin, "The Open Data 500: Putting Research Into Action – The Governance Lab @ NYU." *The GovLab blog*, April 10, 2014, http://thegovlab.org/the-open-data-500-putting-research-into-action/.

[19] For a discussion of the "Bermuda Agreement" that established data-sharing for the Human Genome Project, see Michael Nielsen, *Reinventing Discovery,* (Princeton, NJ: Princeton University Press, 2011), 7.

[20] "HOME – Multiple Myeloma Research Foundation," http://www.themmrf.org/.

Scientists themselves are also finding that sharing data publicly can help them break through logjams in their research. The approach known loosely as crowdsourcing enables thousands of volunteers to help make scientific Open Data more valuable. The website Zooniverse[21] was started by a Ph.D. student who had to look at the structure of galaxies using images from the Hubble telescope. Computers apparently are not as good as humans in doing this kind of image analysis, and the student had to find a way to go through nine hundred thousand images for his work. He and his colleagues came up with this the novel idea of posting the images online, giving simple directions about how to identify the particular type of galaxy they were looking for, and inviting the public to participate. Tens of thousands of people soon volunteered to help. Zooniverse has now applied the same approach to a wide range of scientific problems, including cancer and climate research as well as astrophysics.[22]

### C. *Data from the Crowd*

One of the richest sources of Big, Open Data is social media. There are now three billion tweets sent every week,[23] in addition to countless numbers of opinions written on review websites and elsewhere. All that activity has created a massive amount of data about everything from consumer products and services to political trends. It's now possible to analyze this data through the approach known as sentiment analysis, a set of techniques for analyzing text and other forms of data to extract and analyze opinions and insights about products, services, and trends.[24]

Sentiment analysis is an emerging tool for marketing and business strategy, and can also be used for social good. In Washington, DC over two years, the Mayor's Office has used sentiment analysis to evaluate and analyze what people are saying about government agencies like the Department of Motor Vehicles on social media. As you imagine when they started, people were not saying such great things; they gave

---

[21] "Zooniverse," https://www.zooniverse.org/.

[22] For a discussion of Zooniverse's work, see the interview with Robert Simpson of Oxford in Gurin, *Open Data Now,* 149-151.

[23] "Twitter," https://about.twitter.com/company.

[24] Joel Gurin, "Live-Blogging the Sentiment Analysis Symposium," *OpenDataNow.com blog*, March 6, 2014, http://www.opendatanow.com/2014/03/new-live-blogging-sentiment-analysis-symposium/.

five agencies grades— one got a C plus, the other four got C minuses— but as they have continued to monitor the grades have gotten better and better. The feedback loop has turned out to be a way of improving government.[25]

### D. *Open Data for Smart Cities*

What kinds of data can cities use to learn about and manage themselves? A number of research centers around the world, including the Center for Urban Science and Progress at NYU,[26] are working to analyze all kinds of urban phenomena, ranging from government-collected data to studies of traffic patterns and environmental changes measured by sensors around the city. They believe the data can be used to optimize operations, monitor infrastructure, and improve public health, emergency management, and more.

One of the first applications has been public transportation. City data has been used to develop an app called Nextbus[27] and similar applications that tell you when your bus is coming, so you don't have to wait out in the rain for it. As Nextbus began developing apps for different cities, the cities themselves began preparing transportation data in a standard format to make it easier for the company to provide this service to them.

Another company, called OpenGov,[28] is providing a platform that enables any city in the country to put their financial data in a simple graph form. City budgets have generally been kept in hard-to-read spreadsheets. The OpenGov platform opens the data up in a way that not only leads to greater transparency for citizens, but also makes it possible for city managers to compare their city to neighboring cities on dimensions like financial performance or police overtime.

---

[25] Julie Zauzmer, "Mayor Gray Celebrates D.C.'s Good Grades," *The Washington Post*, July 9, 2013; Gurin, *Open Data Now*, 135-137.

[26] "Center for Urban Science+Progress," http://cusp.nyu.edu/.

[27] "Nextbus," http://cts.cubic.com/en-us/solutions/real-timepassengerinformation /nextbus,inc.aspx.

[28] "OpenGov," https://www.opengov.com/.

## II. OPEN SOLUTIONS TO BIG DATA PROBLEMS

### A. *Making Government Data More Useful*

By one definition, datasets are Big Data if they are characterized by three V's: Volume, Variety, and Velocity. But a fourth V – Value – may be more important than any of these. What good is Big Data if it doesn't help solve a compelling social, business, or scientific problem? And how can we tell which datasets will turn out to have the greatest value?

There's a good argument that you can't predict the value of data until it's released as Open Data, and that all government data should thus be made "open by default" – released as Open Data unless there is a strong privacy, security, or other reason to keep it closed. The Obama Administration, like an increasing number of governments around the world, has tried to make much more government data open. In May 2013, through an Executive Order from the President[29] and a memo from the Office of Management and Budget[30], the administration established the Open Data Policy, stating that all federal data should be made open and easily usable unless there is an express reason not to.

In announcing the Open Data Policy, President Obama said that Open Data is "going to help launch more startups. It's going to help launch more businesses…It's going to help more entrepreneurs come up with products and services that we haven't even imagined yet."[31] The policy was designed specifically to make government data more useful to the private sector, as well as for the public good.

The Open Data Policy requires federal data to be machine readable (meaning that it can easily be used by computers), timely, reusable – all good things. The problem is in getting government data to meet those standards. There are an estimated ten thousand different federal

---

[29] Barack Obama, Executive Order, "Making Open and Machine Readable the New Default for Government Information," May 9, 2013, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

[30] Sylvia M. Burwell et al., Memorandum for the Heads of Executive Departments and Agencies, "Open Data Policy – Managing Information as an Asset," May 9, 2013, http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

[31] Barack Obama, "Remarks by the President at Applied Materials, Inc.-Austin, TX," May 9, 2014, transcript, http://www.whitehouse.gov/the-press-office/2013/05/09/remarks-president-applied-materials-inc-austin-tx.

information systems,[32] and each of those systems has a lot of databases in it. They are legacy systems, they often don't work well technologically, and they are often not interoperable. For example, the Occupational Safety and Health Administration and the Environmental Protection Agency have a number of facilities that they both regulate – OSHA for workplace safety, EPA for pollution. It would be very interesting to mash up those two data sets. But because they operate in different ways, it's now very difficult to do that.[33]

Federal agencies can't practically improve all of their data at once. In order to fulfill the requirements of the Open Data Policy, they may do best to begin with the datasets that have the highest value to the people who use their data.

### B. *Privacy and Personal Data*

It's impossible today to talk about Big Data without addressing concerns about individual privacy. Revelations about the NSA's surveillance program have made many Americans suspicious of data collection overall and of Big Data in particular. From one perspective, the NSA has been an example of what can happen when Big Data is closed to the public – so closed, in fact, that the public does not even know that Big Data collections exist. The NSA has led to calls for greater government transparency and for a rethinking of data collection and privacy rights.[34]

One issue is not only that companies and governments collect Big Data about individuals, but that people don't actually know what's being collected about them. This is where the concept of openness can be helpful. The basic idea is that information about individuals should be "open" to those individuals – that they should know when it's being collected, should be able to download it, and should be able to use it as they see fit.

While we've recently seen the major problems in the way that the Department of Veterans Affairs has handled medical records, it's worth looking at a much more positive initiative that they started

---

[32] Jared Duval, *Next Generation Democracy,* (New York: Bloomsbury, 2010), 174.

[33] Hamid Ouyachi, personal communication to author, May 1, 2014.

[34] See Gurin, *Open Data Now*, Chapter 11, "Privacy, Security, and the Value of Personal Data."

several years ago. The Blue Button program[35] was started to give veterans access to their medical records. When the VA launched it, they thought that only a few thousand people might use it. It soon hit a million users and was then adopted by the private sector. As a result, more than 150 million Americans now have access to their medical records[36] through this kind of application. A similar program, Green Button, now gives Americans data on their own energy usage.[37]

### C. *Creating a See-Through Society*

While the Open Data Policy and other initiatives have focused on making government data open to the public, that's only one aspect of Open Data. We're now seeing a much broader movement to make Big Data of all kinds publicly available, including data from corporations. This trend could help counteract the dark side of Big Data and, in fact, give us a more open and transparent society – what you could call a "see-through society" – than we've seen before.

Some of the pressure for this greater openness is coming from journalistic and advocacy organizations. Pro Publica,[38] the non-profit organization that has developed a very effective, data-driven approach to journalism, is using Big Data that is open to them to do all kinds of investigative reports that would otherwise be impossible. The Sunlight Foundation[39] has done work on Congress and campaign finance, using public data to make these issues transparent to the public.

Another trend is the use of Open Data from consumer complaints to make companies and corporations more publicly accountable and socially responsible. The Consumer Financial Protection Bureau has created a database of complaints[40] on credit cards, mortgages, and

---

[35] "Blue Button Home," *U.S. Department of Veterans Affairs*, last modified March 11, 2014, http://www.va.gov/bluebutton/.

[36] Nick Sinai, "Leading Pharmacies and Retailers Join Blue Button Initiative," *Health IT Buzz*, February 7, 2014, http://www.healthit.gov/buzz-blog/consumer/leading-pharmacies-retailers-join-blue-button-initiative/.

[37] "Green Button," http://greenbuttondata.org/.

[38] "ProPublica, Journalism in the Public Interest," *Pro Publica Inc.,* http://www.propublica.org/.

[39] "The Sunlight Foundation," http://sunlightfoundation.com/.

[40] "Submit a Complaint," *Consumer Finance Protection Bureau*, http://www.consumerfinance.gov/complaint/.

other kinds of financial services. Researchers are analyzing these complaints, rating banks and financial institutions, and seeing those institutions improve their customer service significantly.[41] Other government agencies, including the Consumer Product Safety Commission,[42] are also collecting consumer complaints and making them public. In a similar way, the six federal agencies that handle product recalls are coordinating their data through a single website[43] to bring a new level of information to consumers and increase corporate accountability for product safety.

On a deeper level, investors, advocates, and consumers are now demanding more in-depth information on corporate social responsibility. A company called GoodGuide[44] analyzes about 1500 datasets[45] to give consumers, and companies themselves, insight into the environmental impact and sustainability of a wide range of consumer products. The Carbon Disclosure Project[46] gathers data about the carbon footprint of different corporations and makes it available to institutional investors that collectively manage more than 90 trillion dollars[47] in assets.

The rather vague concept of corporate social responsibility is now being replaced by environmental/social/governance, or ESG, reporting. In this new framework, companies report on specific metrics that describe how their operations affect the environment, local communities, and socially important concerns. Some ESG reporting is voluntary, and some is now being institutionalized at a government level. The Securities and Exchange Commission now requires public companies to report on their use of conflict minerals,

---

[41] Amy Fontinelle, "Why Banks Are Scrambling to Hear Your Complaints," *Forbes*, October 25, 2013, http://www.forbes.com/sites/investopedia/2013/10/25/why-banks-are-scrambling-to-hear-your-complaints/.

[42] "Saferproducts.gov," *Consumer Products Safety Commission*, http://www.saferproducts.gov/.

[43] "Recall.gov," *Consumer Products Safety Commission*, http://www.recalls.gov/.

[44] "GoodGuide," http://www.goodguide.com/.

[45] Interview with GoodGuide founder Dara O'Rourke, April 11, 2013, quoted in Gurin, *Open Data Now*, (New York: McGrawHill, 2014), 103-107.

[46] "Carbon Disclosure Project," https://www.cdp.net/en-US/Pages/HomePage.aspx.

[47] "What We Do," *Carbon Disclosure Project*, https://www.cdp.net/en-US/WhatWeDo/Pages/investors.aspx.

minerals found in many electronic products that are mined under inhumane conditions in the Republic of Congo.[48]

Corporations have been reluctant to release this kind of data about themselves in the past, and it's not likely that they will rush to accept the see-through society overnight. Large companies and their lobbyists are likely to push back on proposed government legislation and advocacy initiatives that call for greater transparency. Some of this resistance is based on the rationale that they could be asked to release proprietary information, while some is framed as an objection to government regulation overall.

The ultimate pressure for transparency, however, may come not from governments or advocates, but from investors. An increasing number of investors are now looking for "sustainable" companies, that is, companies that follow good environmental/social/governance practices that will enable them to succeed over the long haul. For investors, sustainability may be simply an indicator of good corporate governance: Sustainable companies are more likely to be prepared for government environmental mandates, more likely to operate in ways that will not antagonize local governments, and less likely to cause environmental disasters or other scandals that can hurt their valuation.[49]

## III. CONCLUSION: THE DATA FUTURE

Big Data is changing the world for entrepreneurs, businesses, scientists, journalists, and society at large. Now Open Data is having a similar impact. Where Big Data is essentially a technological development, driven by the increased ability to collect data and analyze it, Open Data is more of a philosophical movement, driven by the belief that data should be made available for public use on principle.

The future of data in society will depend largely on how these two concepts develop together. A world where more and more Big Data is collected and held by small, powerful groups could be an Orwellian

---

[48] U.S. Securities and Exchange Commission, "SEC Adopts Rule for Disclosing Use of Conflict Minerals," press release, August 22, 2012, http://www.sec.gov/News/PressRelease/Detail/PressRelease/1365171484002#.VAUeDPldWa8.

[49] For a full discussion of environmental/social/governance measures and Open Data, see Gurin, *Open Data Now*, Chapter Six, "Green Investing: Betting on Sustainability Data."

future. But if Big Data becomes open as well, the risks will be smaller and the potential benefits bigger.